

**HOUSING: PRICE PREDICTION**

Submitted by:

Kunal Chand

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to my teacher

Dr. Deepika Sharma from **Data trained** as well as **Flip Robo**, who gave me the golden opportunity to do this wonderful project on the topic (**HOUSING: PRICE PREDICTION**), which also helped me in doing a lot of research and I came to know about so many new things I am thankful to them.Secondly, I would also like to thank my parents who helped me a lot in finalizing this project within the limited time frame.

**INTRODUCTION**

* Business Problem Framing

The company is looking at prospective properties to buy houses to enter the market. Using Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not.

* Conceptual Background of the Domain Problem

The main problem is found variables which impact most on the price of the house, also the features which are nowadays so important to live a healthy and prosperous life.

* Review of Literature

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data, and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can analyze data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are, and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working. We divided the data 8:2 for Training and Testing purposes respectively.

* Motivation for the Problem Undertaken

Every problem of Machine learning gives us chance to enhance and develop problem-solving skills. These Problems do’s the same.

When this real-life problem of predicting the future pricing of houses, whether to enter the market or not and with help of A. I technology affordable houses for the future generations is in under development.

As Data scientists it is our role to help companies to understand the market better with older data we have for constructing the houses according to that only and make profitable models.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

As for any basic model building, we have to understand the type of target variable, the data of the target variable is continued or classified.

Data Analysis is always the difficult part, for better understanding different kinds of bar plots, distribution plots are created with the target Column for finding the insights of the dataset we have.

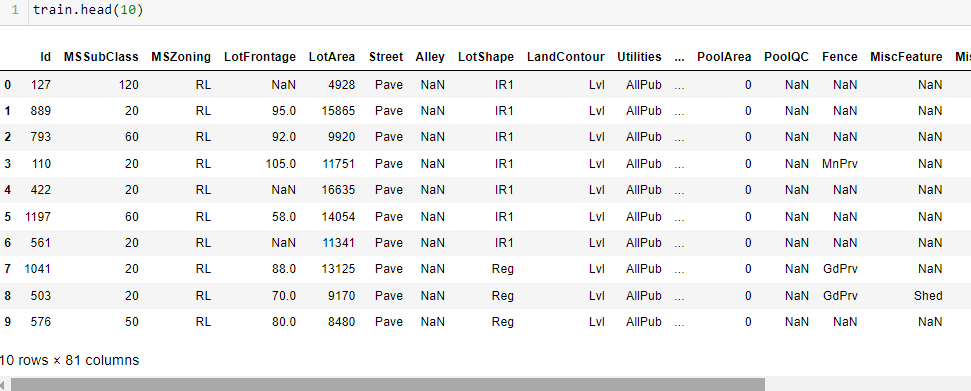
Analytical Modelling always starts with the target variable we have, and in that case, our target variable is Sales Price, for that, we create some box plots with the target variable to understand which feature columns help to learn the model best and which feature columns reduce the accuracy of the model.

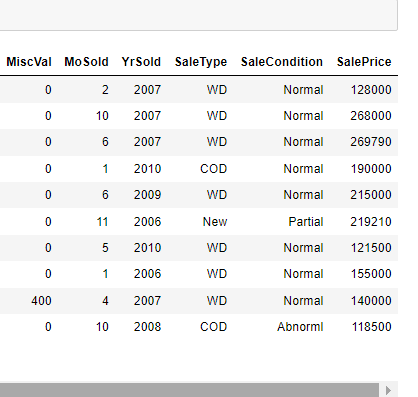
And after finding the relation and correlation with the target variable we choose either Regression Model or Classification Model. Here in this problem, our target feature column continues so we build our Machine Learning model on Regression.

* Data Sources and their formats

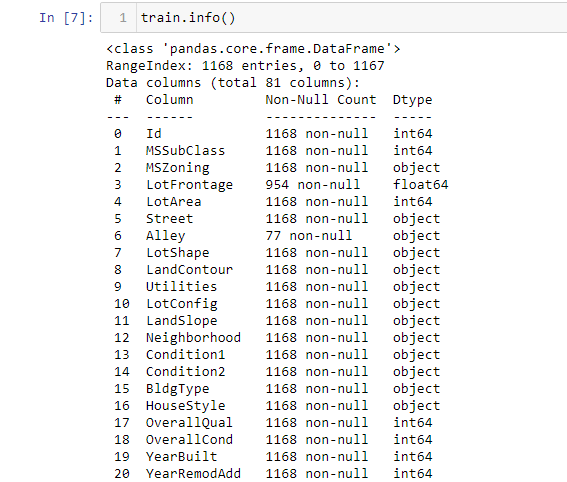
A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

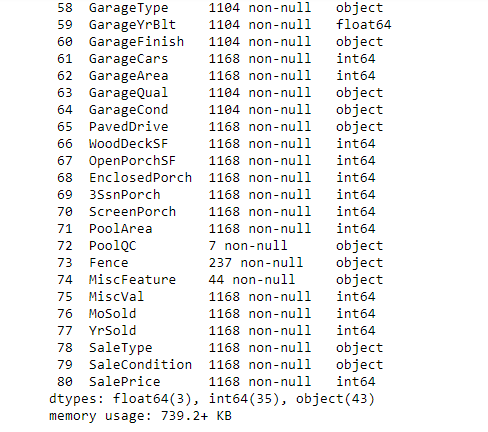
**Dataset looks as follows-**

****

****

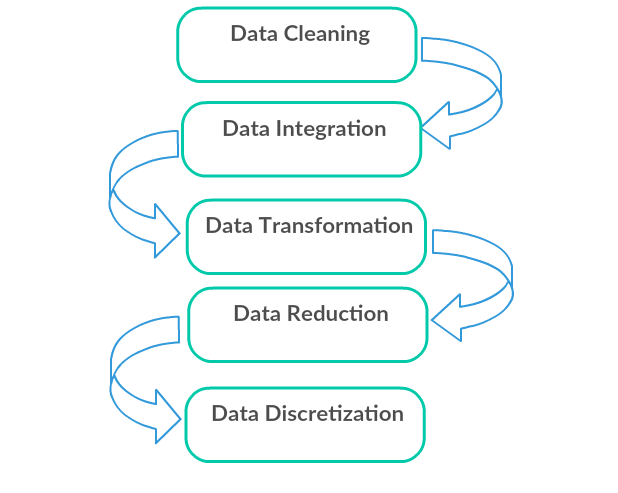
**Dataset Information looks as follows-**

****

****

* Data Pre-processing Done

Data pre-processing can refer to the manipulation or dropping of data before it is used to ensure or enhance performance, and is an important step in the data mining process.



1. Data Cleaning: First we clean the data which have no use in prediction like the ID column, then we drop the data which has a high no of missing percentages.
2. Data Integration: then we do some EDA process for finding out the meaning full insights of the data.
3. Data transformation is the process of changing the format, structure, or values of data; we use a labeled encoder for coding the object data into integer data.
4. Data Reduction: it is the process of finding the most correlated columns, and combining them because the machine does not understand which feature columns impact the most on accuracy.
5. Data discretization converts a large number of data values into smaller once, so that data evaluation and data management becomes very easy, using box plots is makes a clear understanding of the data.

Data Inputs- Logic- Output Relationships

Finding the relation between the columns of the input with the target column is always important.

**1. Overall Quality vs. Sales Price**

Observations:

1. As from the above observations, Overall quality increases the Sale price of the House.

2. Overall quality will depend on other features as well.

3. As the Score increases Sales price also increases.

**2. Year Built vs. Sales Price**

Observations:

1. As from above observations, as we years increases Sales Prices also increases.

2. Property Rates increase with time, as years pass house prices will increase gradually.AS expected.

**3. Sales Condition vs. Sales Price**

Observations:

1. As from above observations and plotting we can easily see the Sales Conditions increases the Sales Pricing.

2. Abnormal condition increases the sales Prices as expected.

**4. Lot Size vs. Sales Price**

Observations:

1. As from the above observations and plotting we can easily see the Lot Area increases the Sales Pricing.

* State the set of assumptions (if any) related to the problem under consideration

During Data cleaning we assume that the columns having more than 50% Nan values are present are not affecting the accuracy of the model.

* Hardware and Software Requirements and Tools Used

**Python**

Python is widely used in scientific and numeric computing:

SciPy is a collection of packages for mathematics, science, and engineering.

Pandas are data analysis and modeling libraries.

Libraries Used for this Project include –

1. Pandas

2. NumPy

3. Matplotlib

4. Seaborn

5. Scikit Learn

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

After analyzing the dataset, I observe that many of the feature columns are object type so first, we have to convert them in integer or float type so that machine interpret the data and for that we do label encode all the feature column.

After label encoding, we find that many feature columns have Nan values so we use mean and median for filling that missing data,

Then find the correlation between the columns with target columns and delete the non-related feature columns.

We observe that the target column is skewed so we remove the skewness of the target column because normal data gives better results when we make the M.L model.

The target column is continuous type so we start work on Regression models building.

* Testing of Identified Approaches (Algorithms)

1. Linear Regression
2. Regurgitation:

Lasso & Ridge Regression

1. Ensemble techniques

Decision Tree Regression

Random forest Regression

1. Gradient Boosting Regression
2. Support vector machine
3. K-nearest Neighbour Regression

* Run and Evaluate selected model

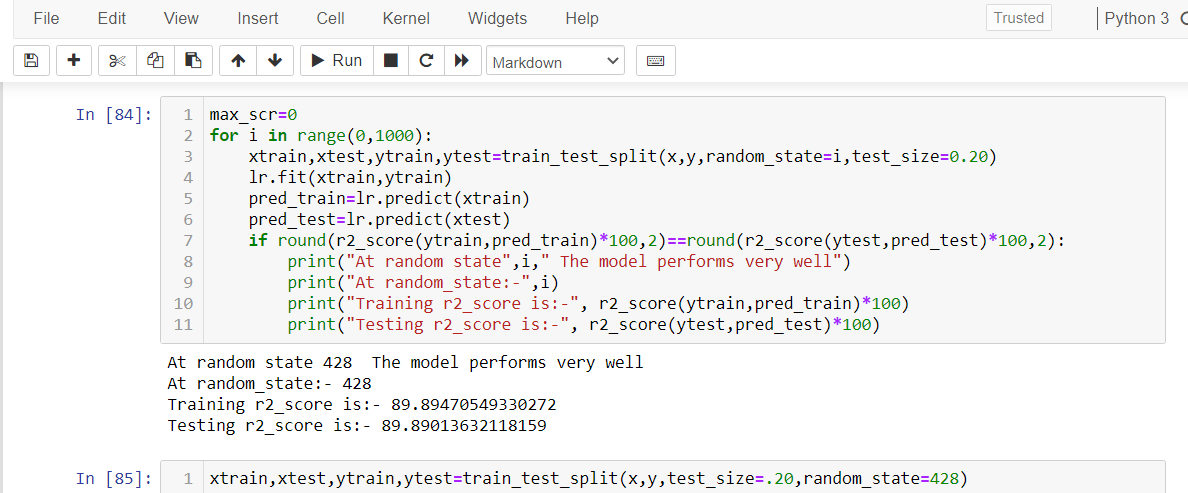
MODELS USED

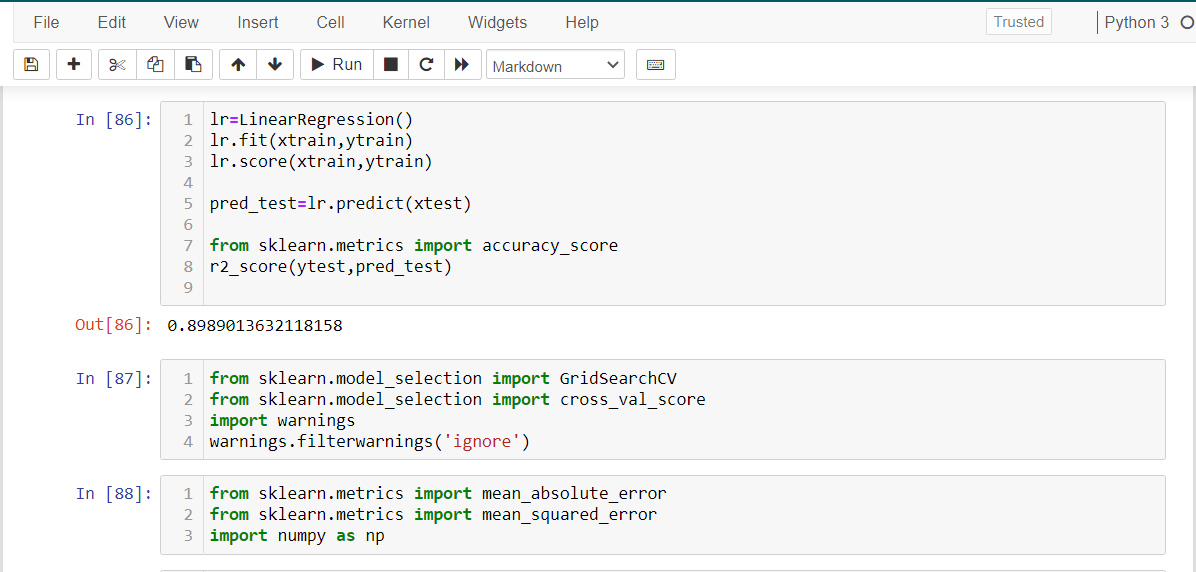
**Linear Regression Model**

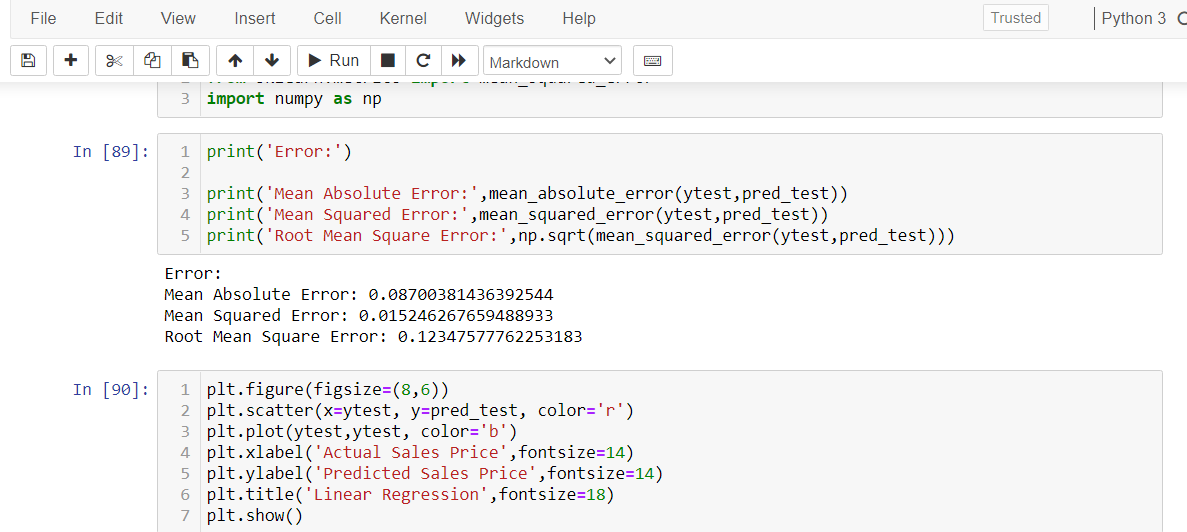
• Linear Regression is a machine learning algorithm based on supervised learning.

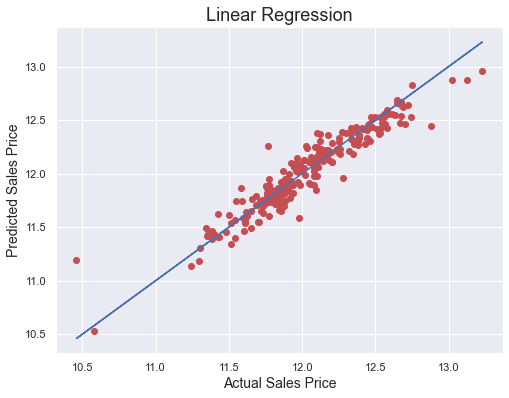
• It performs a regression task. Regression models a target prediction value based on independent variables.

• It is mostly used for finding out the relationship between variables and forecasting.









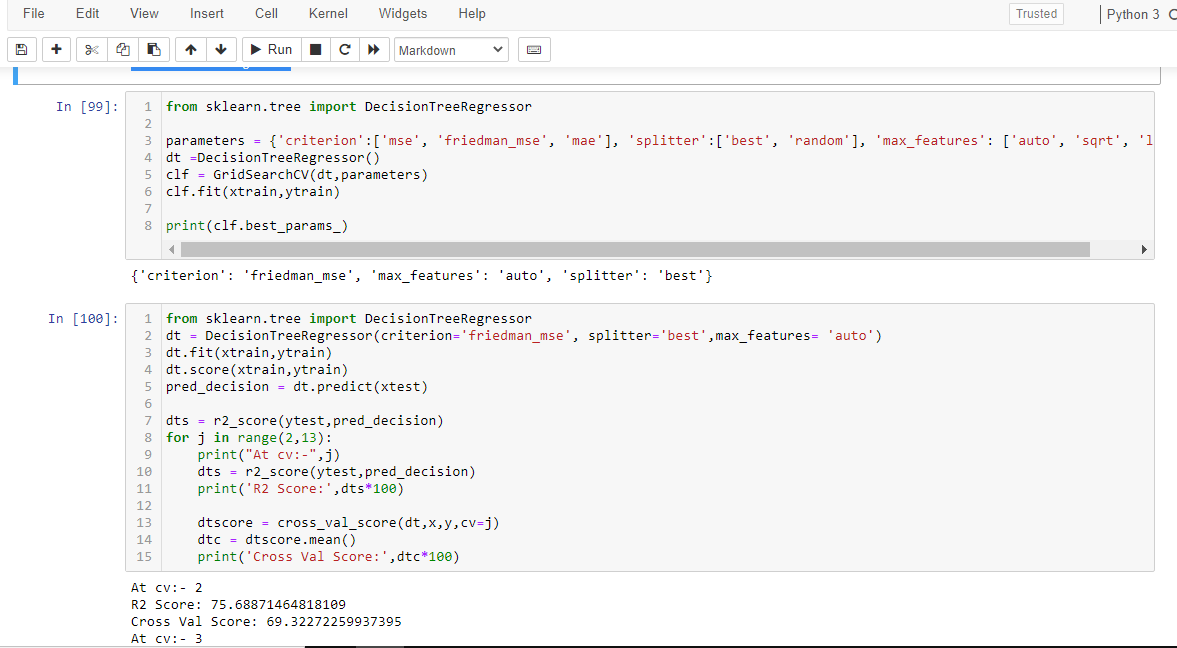
**Observations:**

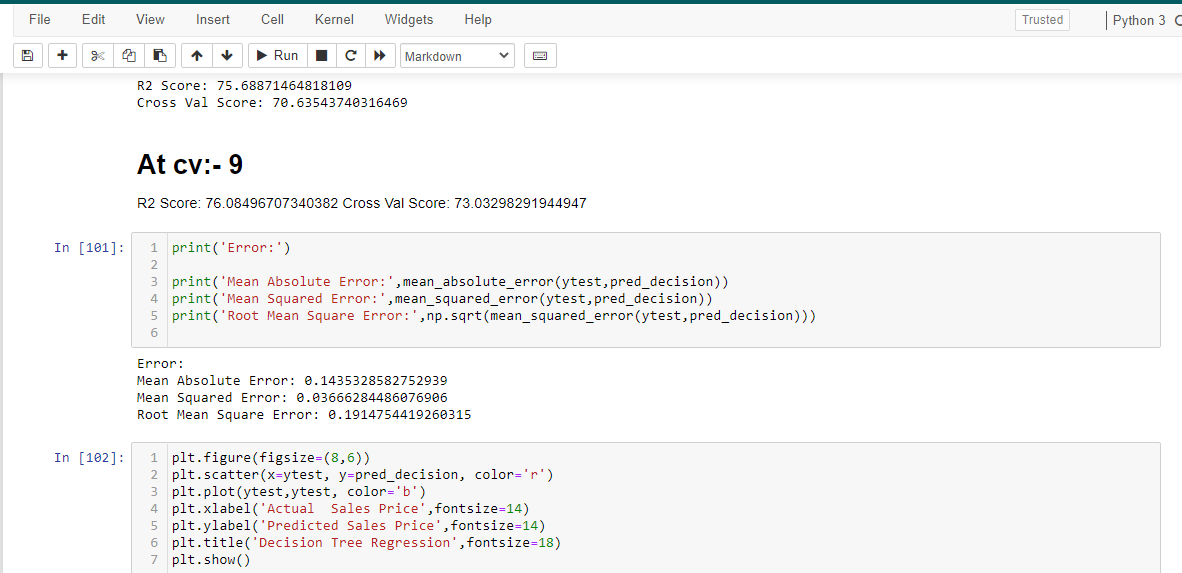
1. This Linear Regression Performs with 90% accuracy for predicting house prices.
2. We use the best-fit line and we can easily see that most of the price points are fall on the line.

# Ensemble Techniques:

Decision Tree Regression

A decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast, and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree that corresponds to the best predictor is called the **root node**. Decision trees can handle both categorical and numerical data.







**Observations:**

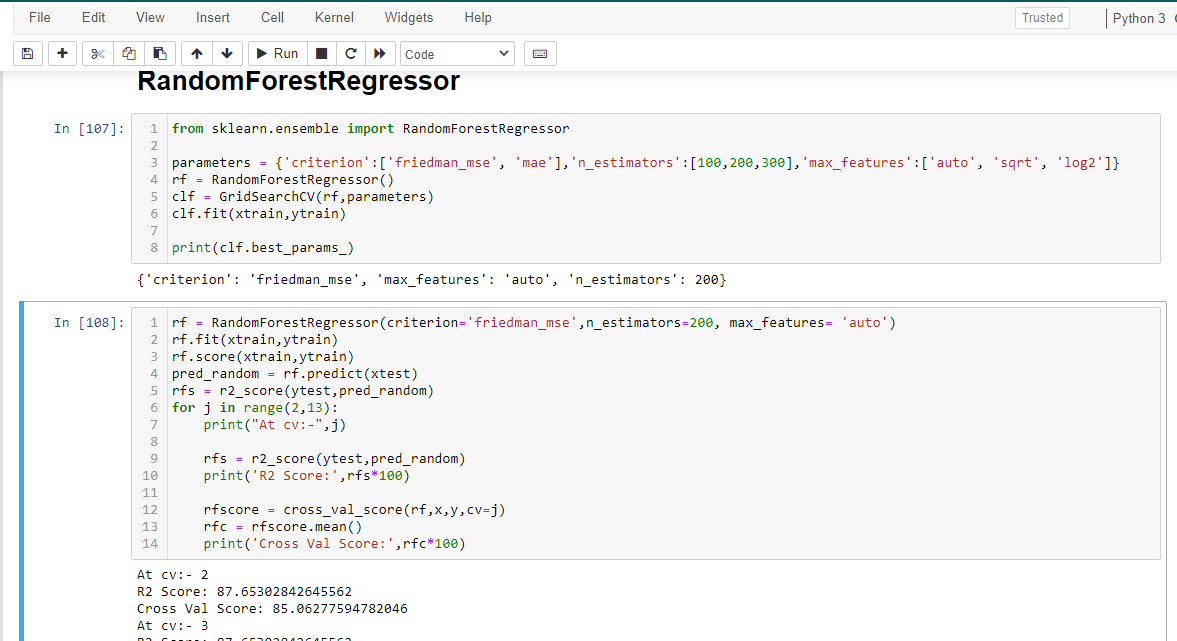
1. This Decision Tree Regression Performs with 76% accuracy for predicting house prices.
2. After predicting and plotting the predicted data on the best fit line we observe that DTR is not so accurate.

**Random Forest Regression Model**

1. A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging.

2. Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement.

3. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.







Observations:

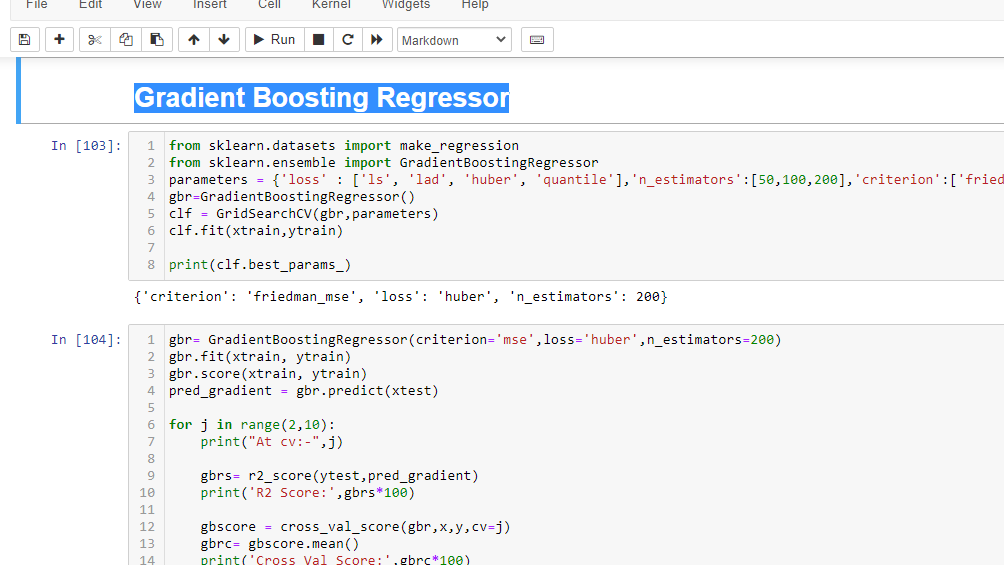
1. RFR performs well but not that well.

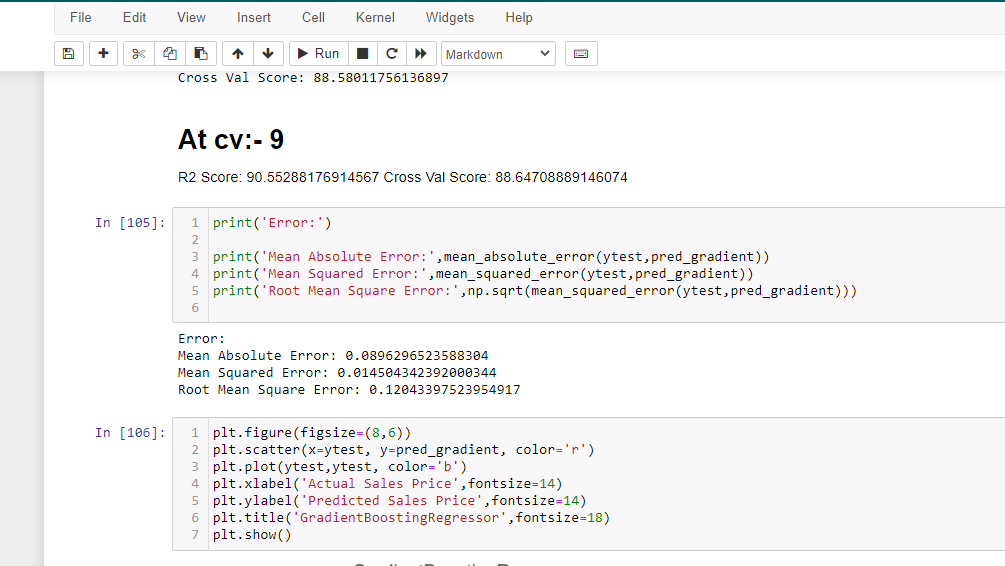
2. CV at 6 is giving good results.

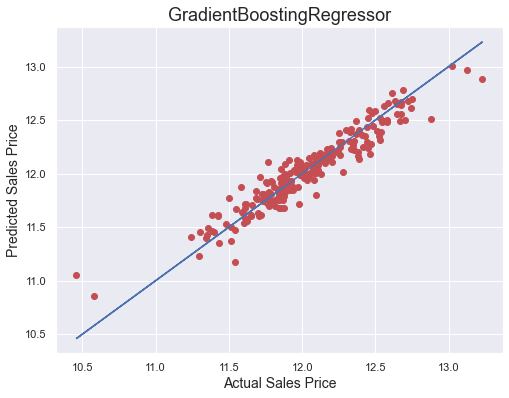
3. After predicting and plotting the predicted data on the best fit line we observe that RFR is not so accurate.

# Gradient Boosting Regression

Gradient boosting is a machine learning technique for regression, classification, and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is a weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest. It builds the model in a stage-wise fashion as other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.







Observations:

1. GBR performs better and gives optimum results.
2. After predicting and plotting the predicted data on the best fit line we observe that GBR is accurate.

# Support Vector Regression

# SVMs or Support Vector Machines are one of the most popular and widely used algorithms for dealing with classification problems in machine learning. However, the use of SVMs in regression is not very well documented. This algorithm acknowledges the presence of non-linearity in the data and provides a proficient prediction model.

# Screenshot (1190).png

# Screenshot (1191).png

# download (4).png

Observations:

1. SVR performs better and gives optimum results.

2. After predicting and plotting the predicted data on the best fit line we observe that SVR is accurate.

3. But when we observe that cv is not better than GBR.

# K-nearest Neighbors Regression

KNN regression is a non-parametric method that, intuitively, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood. The size of the neighborhood needs to be set by the analyst or can be chosen using cross-validation (we will see this later) to select the size that minimizes the mean-squared error.

While the method is quite appealing, it quickly becomes impractical when the dimension increases, i.e., when there are many independent variables.

# Screenshot (1195).png

# Screenshot (1196).png

# download (5).png

Observations:

1. KNN performs not well and gives no proper results.

2. After predicting and plotting the predicted data on the best fit line we observe that KNN is far behind from remaining algorithms.

* Key Metrics for success in solving a problem under consideration

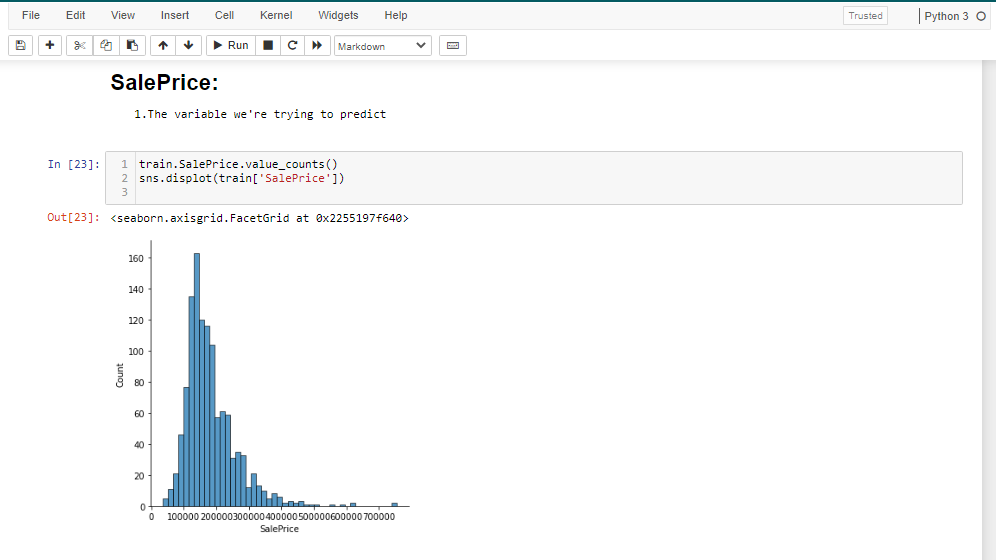
What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

* Visualizations

Data visualization is the graphical representation of information and data. By using charts, plots, and graphs data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Starts from Sales Price analysis.



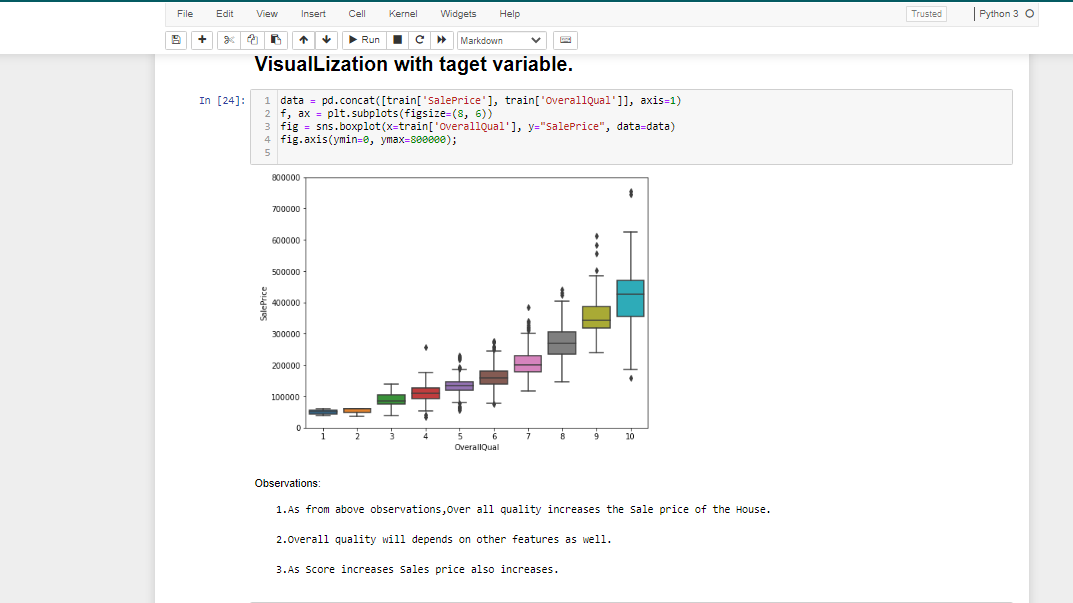
**Observations:**

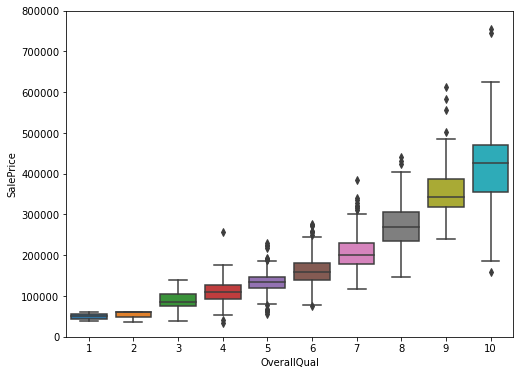
1. As seen from the above distribution plot that Sale Prices are not equally distributed.

2. Data is skewed towards the right.

3. Major Sale Prices are distributed between 100000-400000 dollars.

**Sales Price and overall quality**





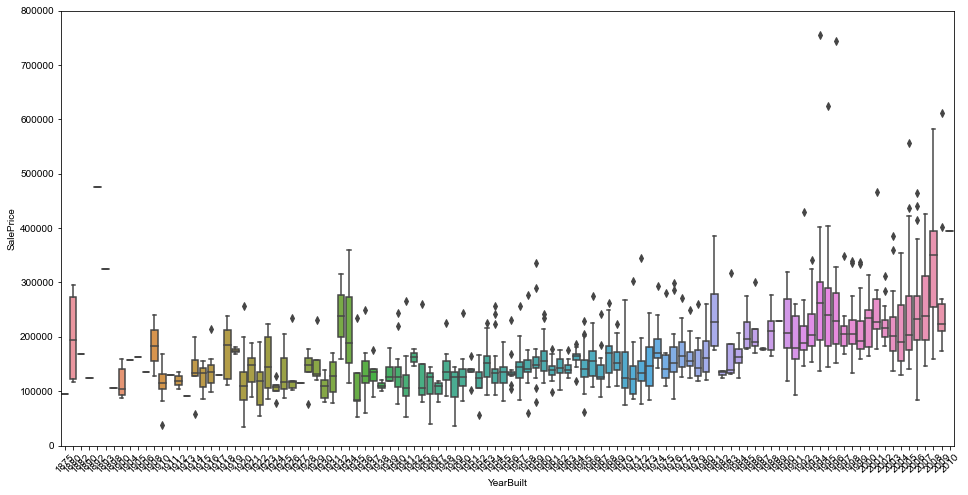
Observations:

1. As from the above observations, Overall quality increases the sales price of the house.

1. The overall quality will depend on other features as well.

3. As the Score increases Sales price also increases.

**Sales Price and built year.**

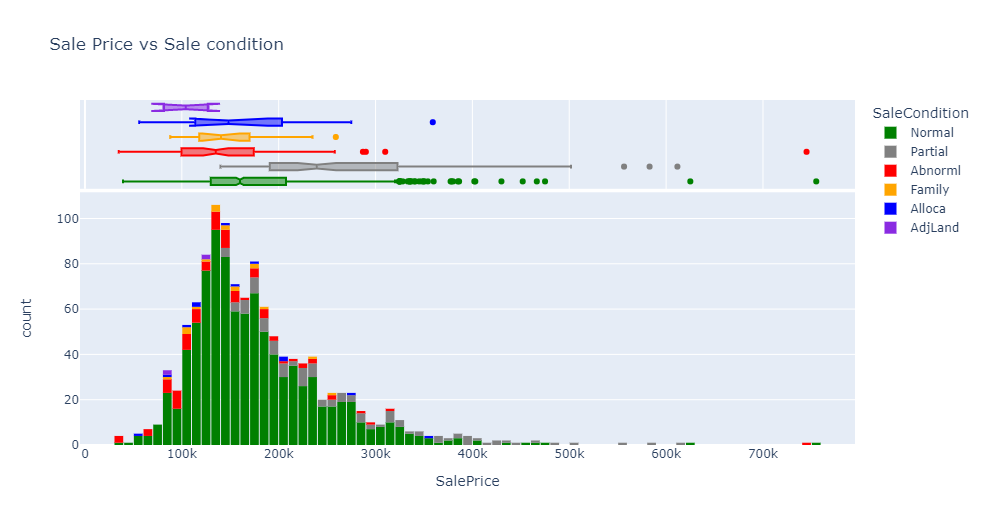


**Observations:**

1. As from above observations, as we years increases Sales Prices also increases.

2. Property Rates increase with time as the property gets old there prices are getting higher.

**Sales Price vs. Sales Condition**



Observations:

1. As from above observations and plotting we can easily see the Sales Conditions increases the Sales Pricing.

2. Abnormal condition increases the sales Prices as expected.

**Sales Price vs. Total Basement Area**



Observations:

1. As from the above observations and plotting we can easily see the total basement area increases the Sales Pricing.

**Sales Price vs. Lot Area**



Observations:

1. As from the above observations and plotting we can easily see the Lot Area increases the Sales Pricing.

**Sales Price vs.** **Ground Live Area**

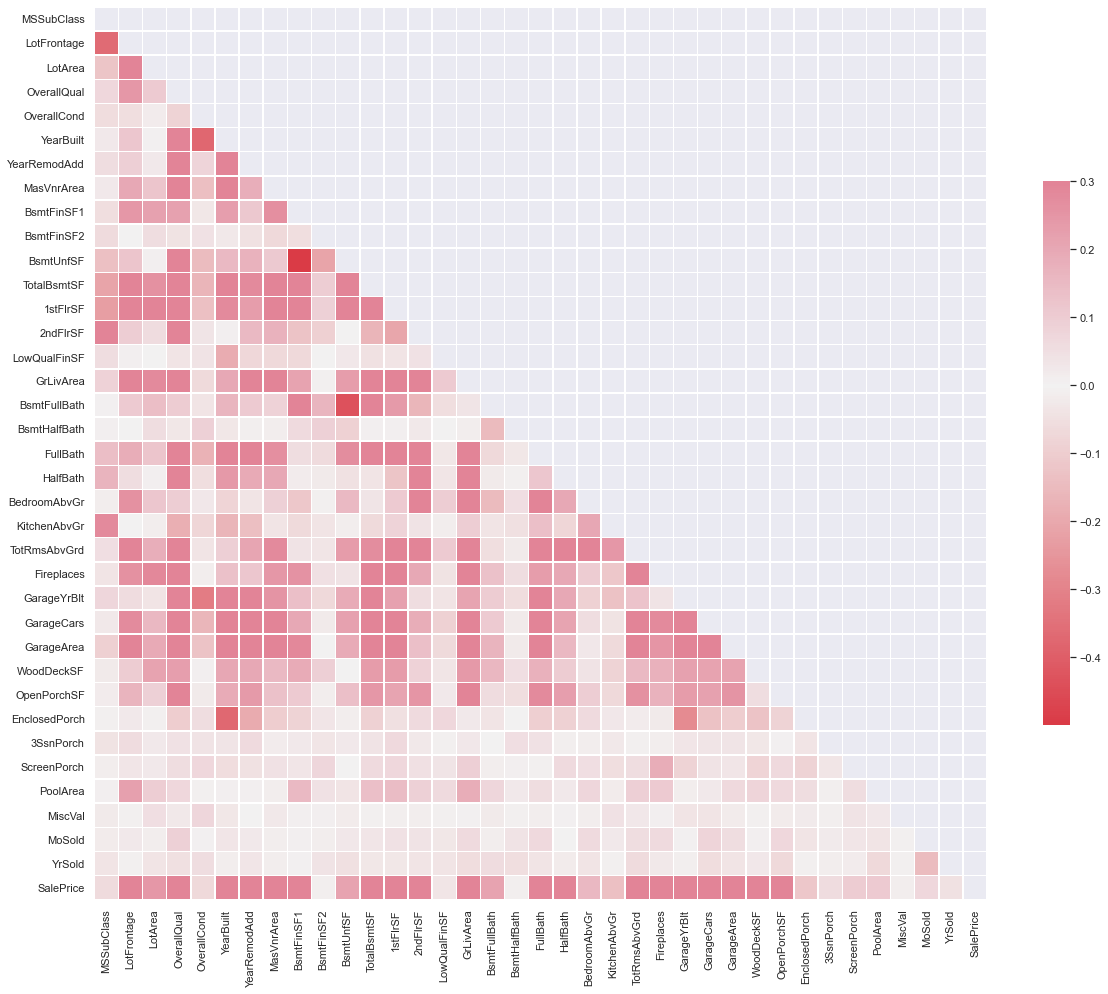


Ground Live Area: Above grade (ground) living area square feet

**Observations:**

1. As from the above observations and plotting we can easily see the Ground Live Area increases the Sales Pricing

**CORRELATION BETWEEN THE COLUMNS:**

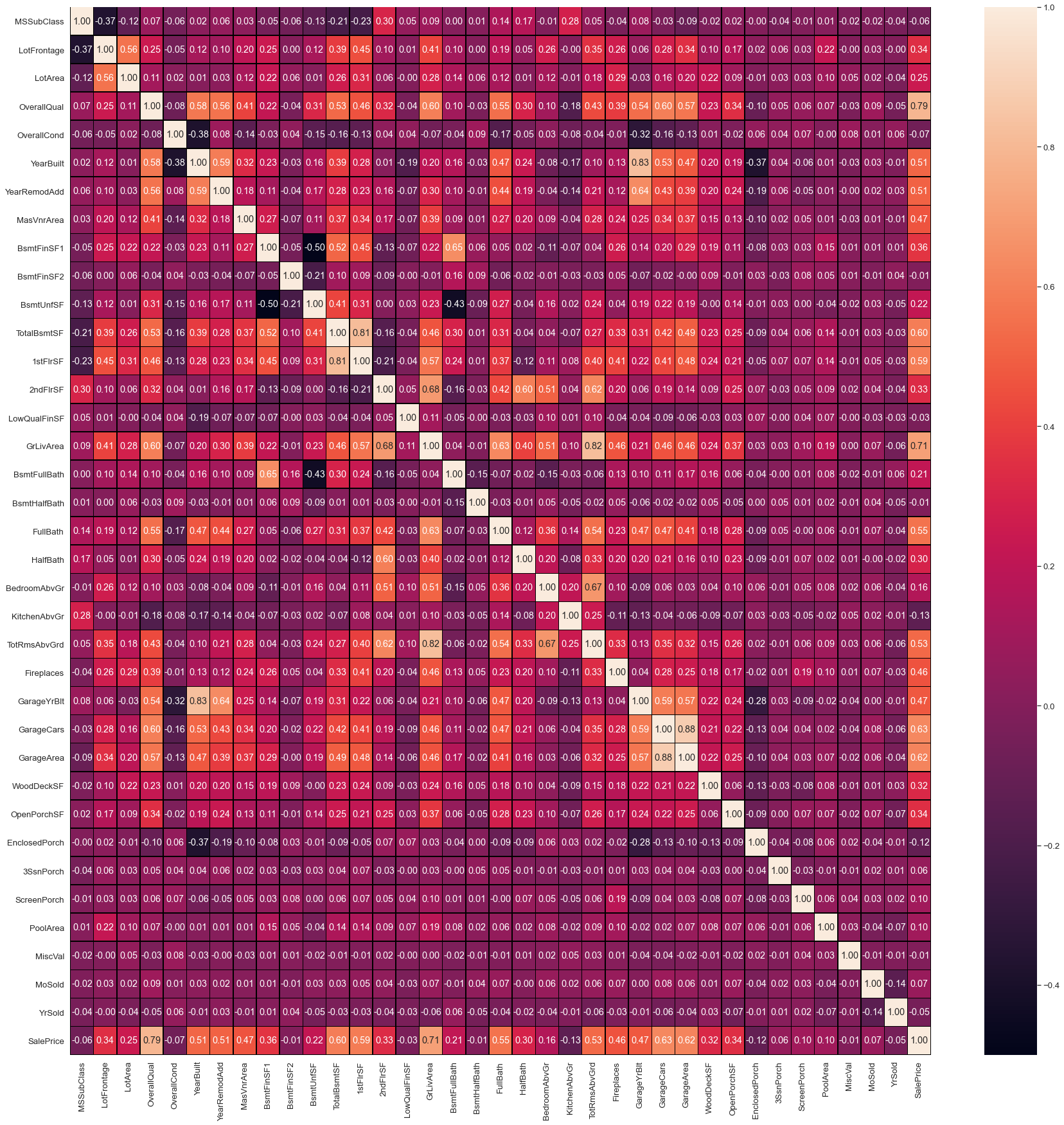


**Observations:**

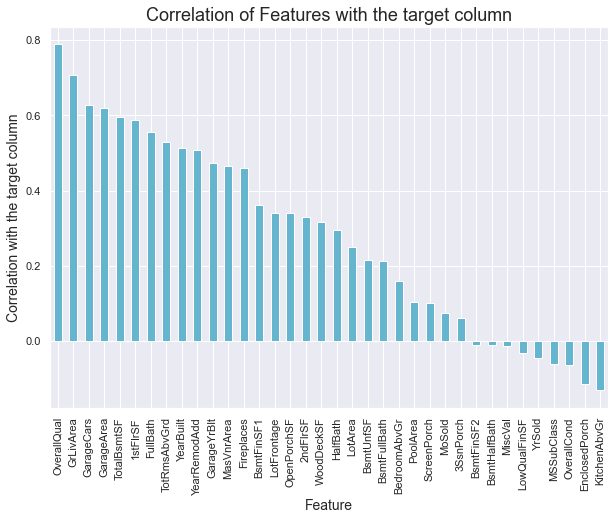
1. Light color boxes are correlated with target columns.

2. Dark color boxes are negatively correlated with target columns.

3. Same color boxes having multi collinear.

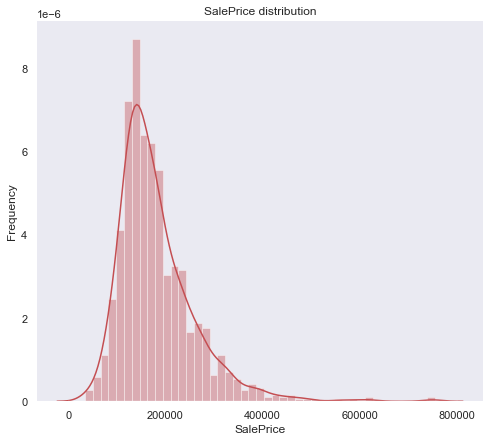






# Check the skewness:

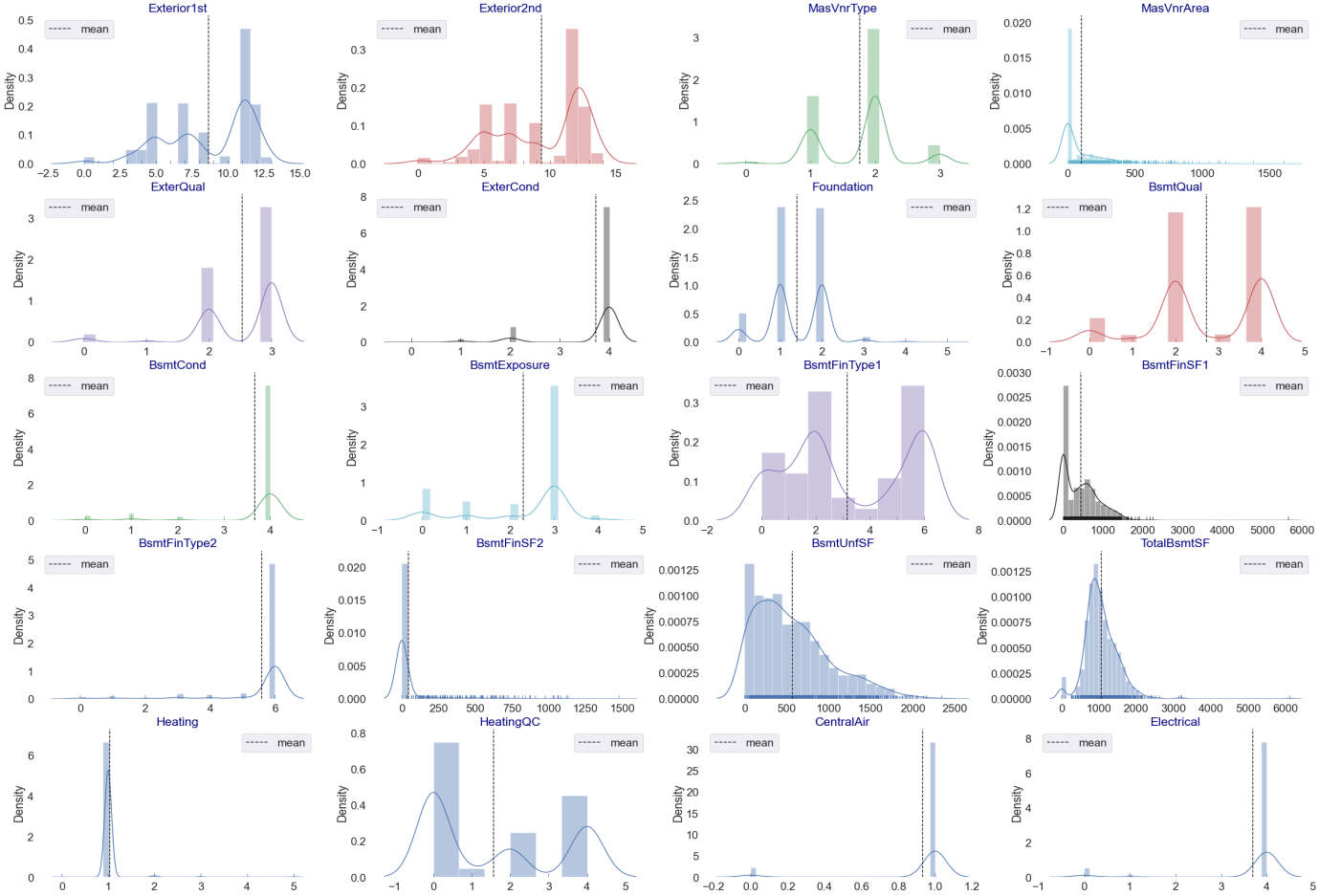
# Now use subplot and distribution plot to check data are normalized or not.

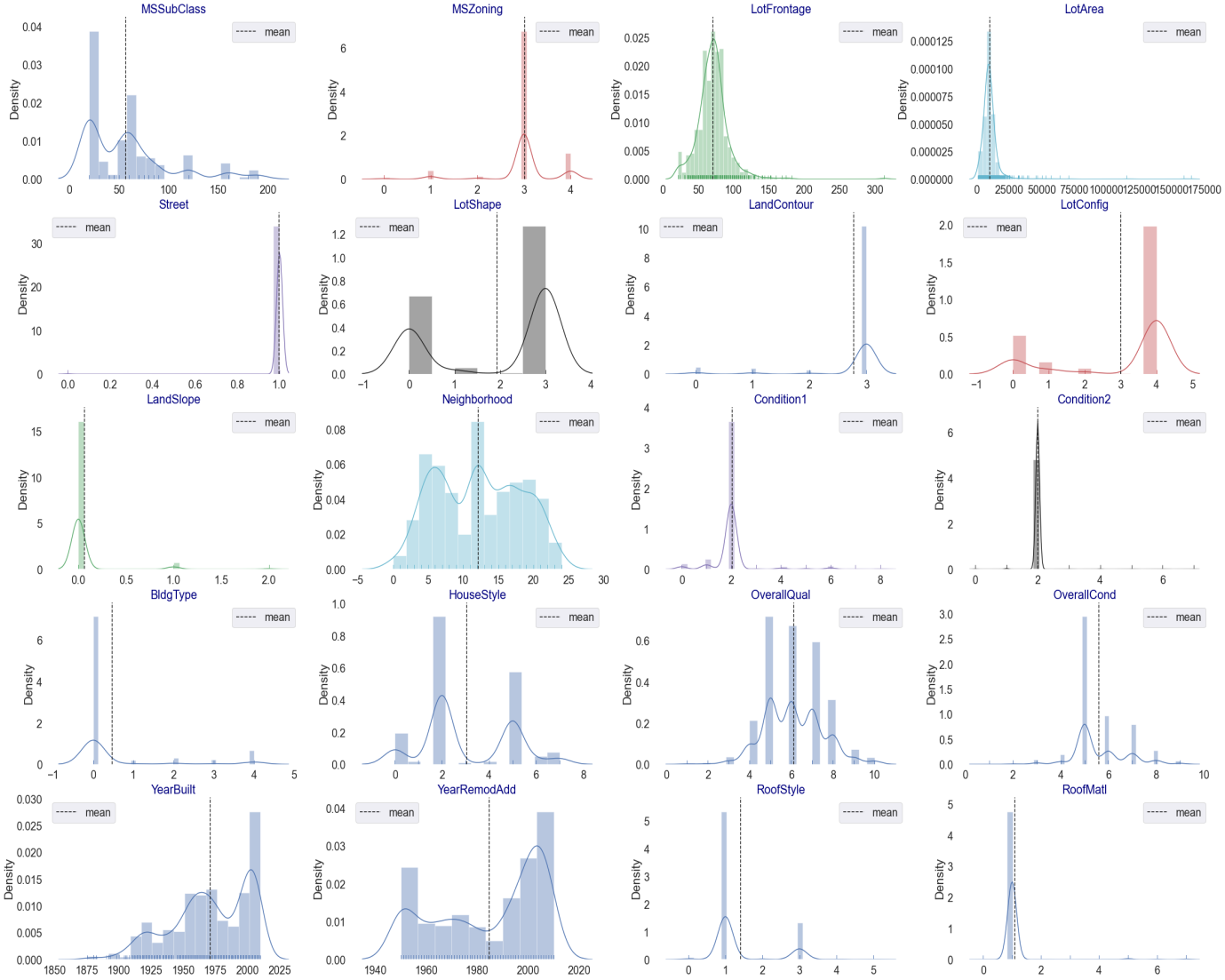
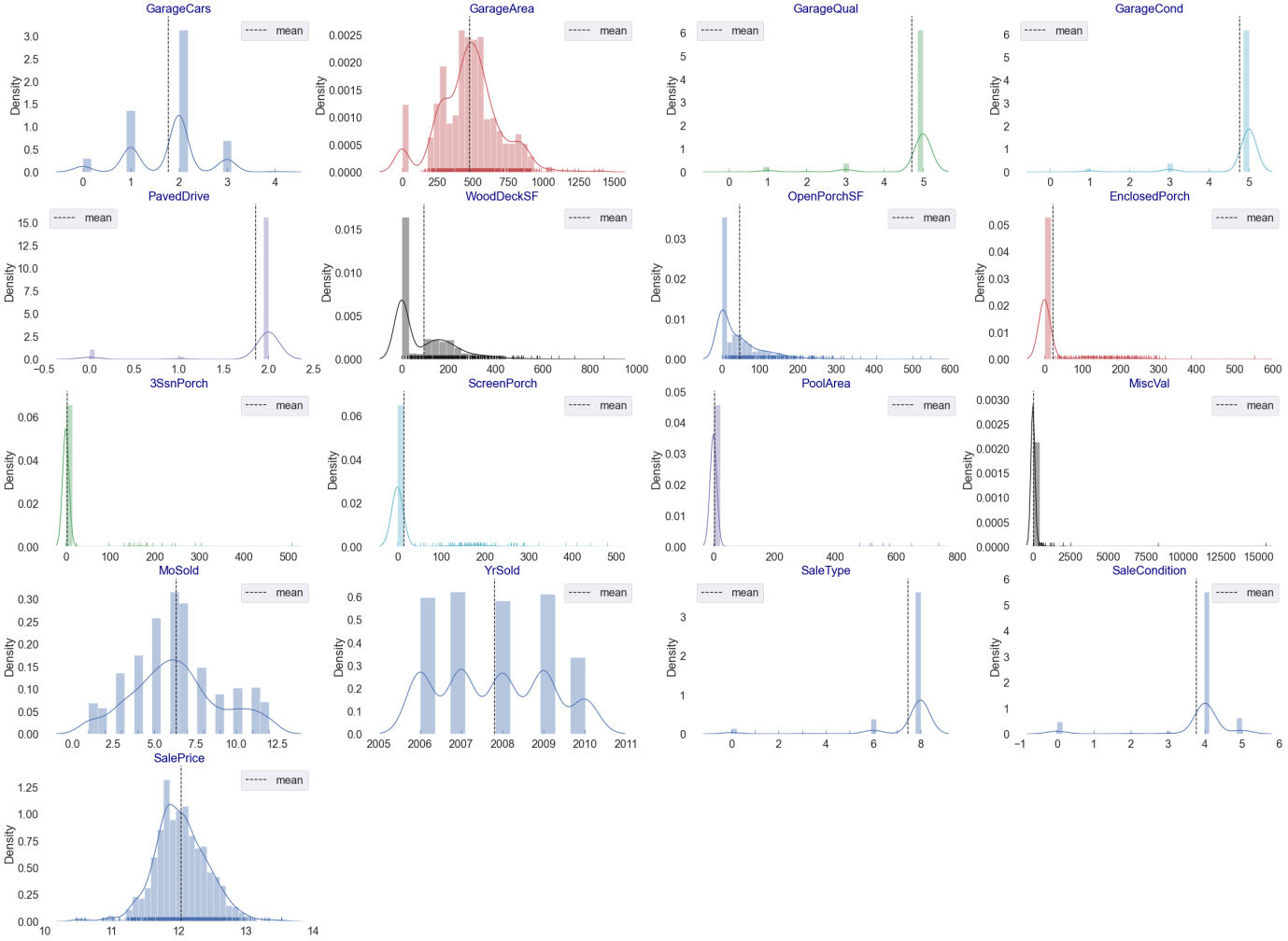


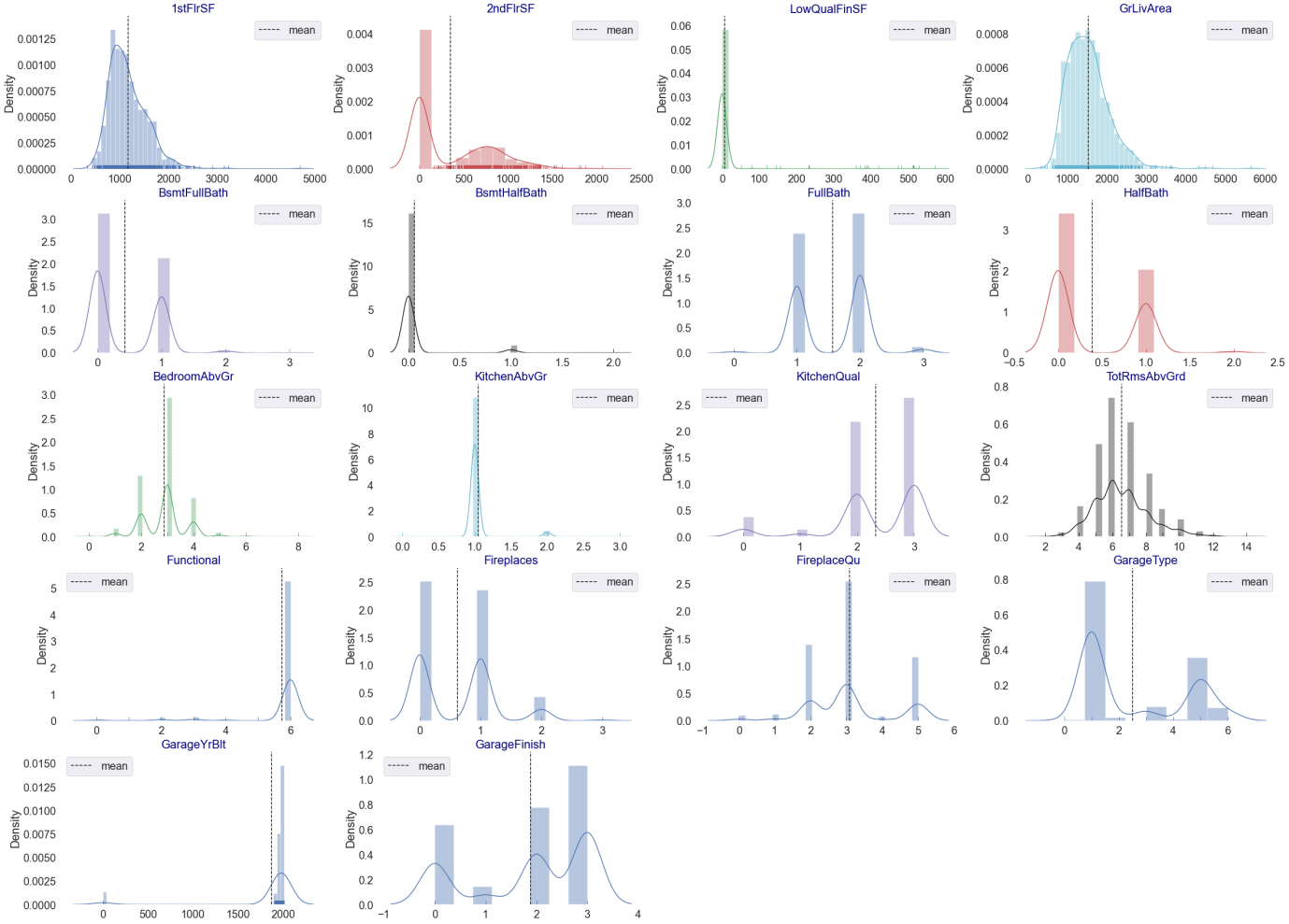
Observations:

1. The Sale Price is skewed to the right. This is a problem because most ML models don't do well with non-normally distributed data.

**The skewness of all the Feature Columns**



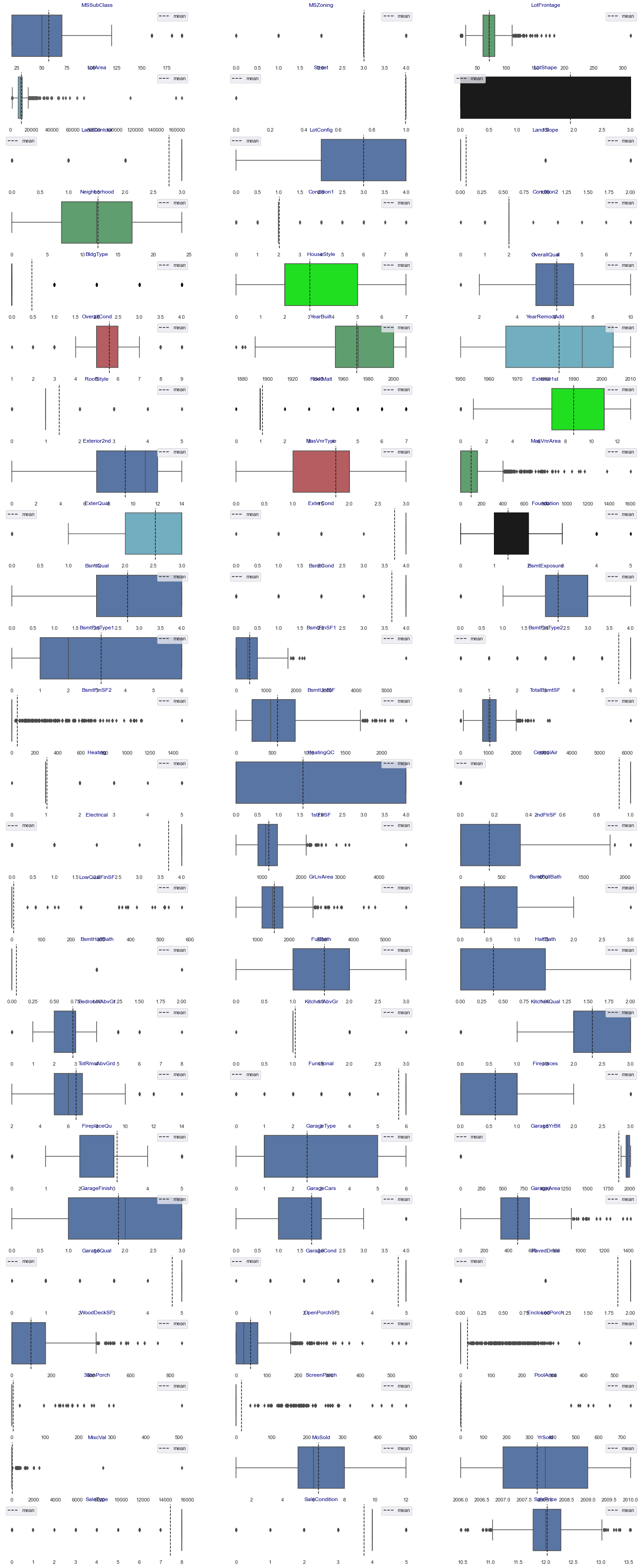




**observations:**

1. From the above plotting of distribution plot, we see that some features columns are not normally distributed.
2. Some columns are skewed towards the right.
3. Building blocks are out of the normal curve hence outliers are present.

**Detecting outliers:**



**Observations:**

1. From the above plotting we see that outliers are present in multiple features columns.

* Interpretation of the Results

The best model is Gradient Boosting Regression. Since the difference between the percentages score of cross-validation and R2\_score is optimum.

At CV: - 9

R2 Score: 90.55288176914567

Cross Val Score: 88.64708889146074

**CONCLUSION**

* Key Findings and Conclusions of the Study

So, our Aim is achieved as we have successfully ticked all our parameters as mentioned in our Aim Column. It is seen overall Quality is the most effective attribute in predicting the house price and that the Gradient Boosting Regression is the most effective model for our Dataset with an R2 score is 0.9055.

* Learning Outcomes of the Study in respect of Data Science

That's it! We reached the end of our exercise.

Throughout this kernel, we put in practice many of the strategies for predicting the prices of the house. We philosophized about the variables, we analyzed 'Sale Price' alone and with the most correlated variables, we dealt with missing data and outliers, we tested some of the fundamental statistical assumptions and we even transformed categorical variables into dummy variables. That's a lot of work that Python helped us make easier.

* Limitations of this work and Scope for Future Work

Limitations of this work are as follows:

1. This study works only in the Australian market.
2. Many Feature columns are related to big cities not working in small cities houses or tear 3 and tear 4 cities and towns.
3. This study did not work in Asian Market.

These are some basic limitations of this study.

For future work, we need data from Asian houses as well or mixed data globally, which predicts house prices well.

Also, the accuracy of predicting is not 100% so many more models to test and which predict with 100% accuracy.

That’s all from this Project Report.

**Thank you**

Data exploration is the first step in data analysis and typically involves summarizing the main

characteristics of a data set, including its size, accuracy, initial patterns in the data, and other

attributes. It is commonly conducted by data analysts using visual analytics tools, but it can

also be done in more advanced statistical software, Python. Before it can analyze

data collected by multiple data sources and stored in data warehouses, an organization must

know how many cases are in a data set, what variables are included, how many missing

values there are, and what general hypotheses the data is likely to support. An initial

exploration of the data set can help answer these questions by familiarizing analysts with the

data with which they are working.

We divided the data 9:1 for Training and Testing purposes respectively.

Data exploration is the first step in data analysis and typically involves summarizing the main

characteristics of a data set, including its size, accuracy, initial patterns in the data and other

attributes. It is commonly conducted by data analysts using visual analytics tools, but it can

also be done in more advanced statistical software, Python. Before it can conduct analysis on

data collected by multiple data sources and stored in data warehouses, an organization must

know how many cases are in a data set, what variables are included, how many missing

values there are and what general hypotheses the data is likely to support. An initial

exploration of the data set can help answer these questions by familiarizing analysts with the

data with which they are working.

We divided the data 9:1 for Training and Testing purpose respectively.